Server I/O Performance Measurement and Analysis

ESG Server Architecture Lab
Intel Corporation

WinHEC
March 27, 1998

Agenda

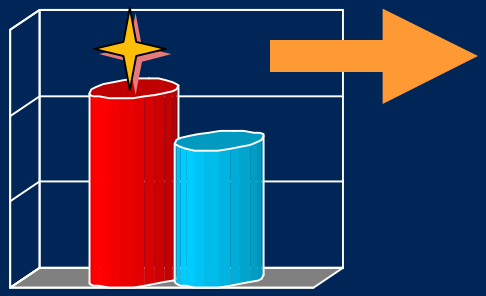
- Why Measurement and Analysis?
- The Science of Measurement
- Analysis Insights
- Summary

Why Measurement & Analysis?

Need to See the Whole Story

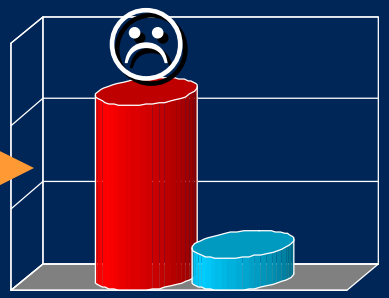
400% ***worse*** PCI utilization!

25% ***better*** throughput...



MBs per Second

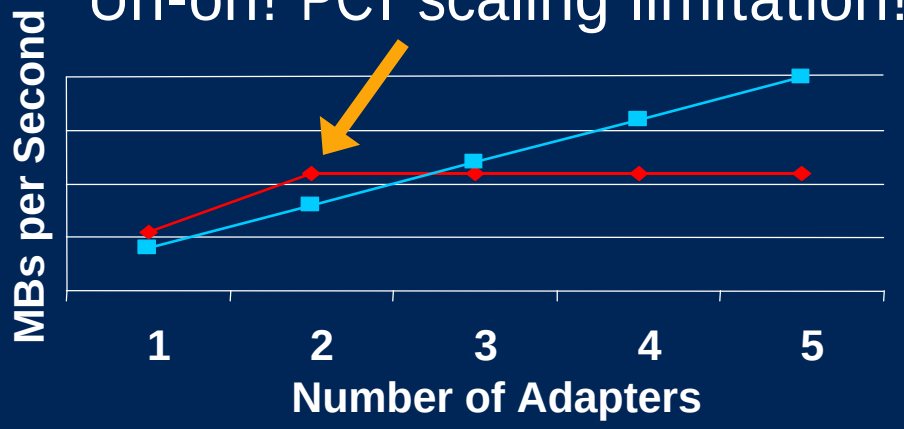
But Wait!



PCI Bus Utilization

Remember: good performance does not imply good scalability!

Uh-oh! PCI scaling limitation!



Data from Intel research

Why Measurement & Analysis?
I/O Efficiency and Scalability is the Key

- I/O is growing rapidly
- I/O is absolutely critical to NT server performance
- I/O will bottleneck if implemented carelessly
- Certain disk subsystems “do more with less”
  - ➔ “Disks == Dollar$”
- I/O efficiency is even more crucial for NICs
  - ➔ small packets, verbose control
- High Volume Servers + NT = Success,
  if I/O is efficient *and* scalable!

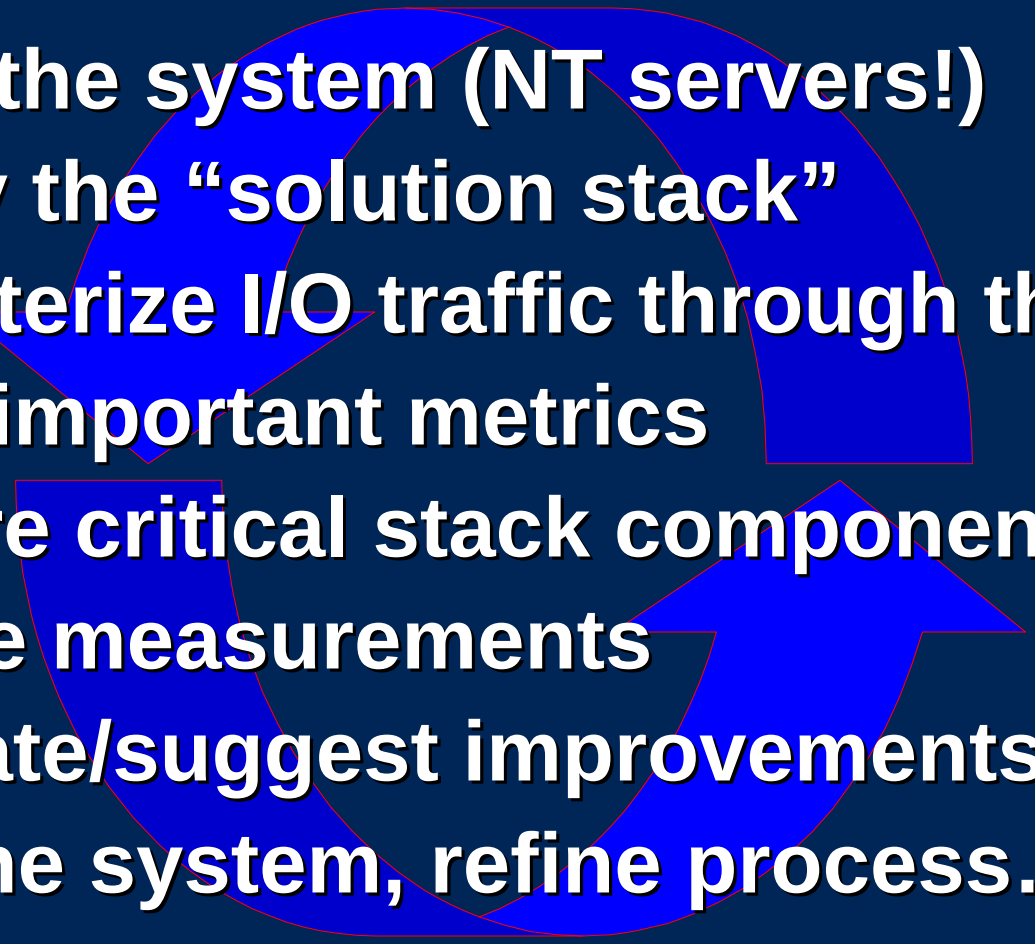
Measurement and analysis can drive dramatic improvements in next generation designs

Agenda

- Why Measurement and Analysis?
- The Science of Measurement
- Analysis Insights
- Summary

The Science of Measurement

The Basic Process

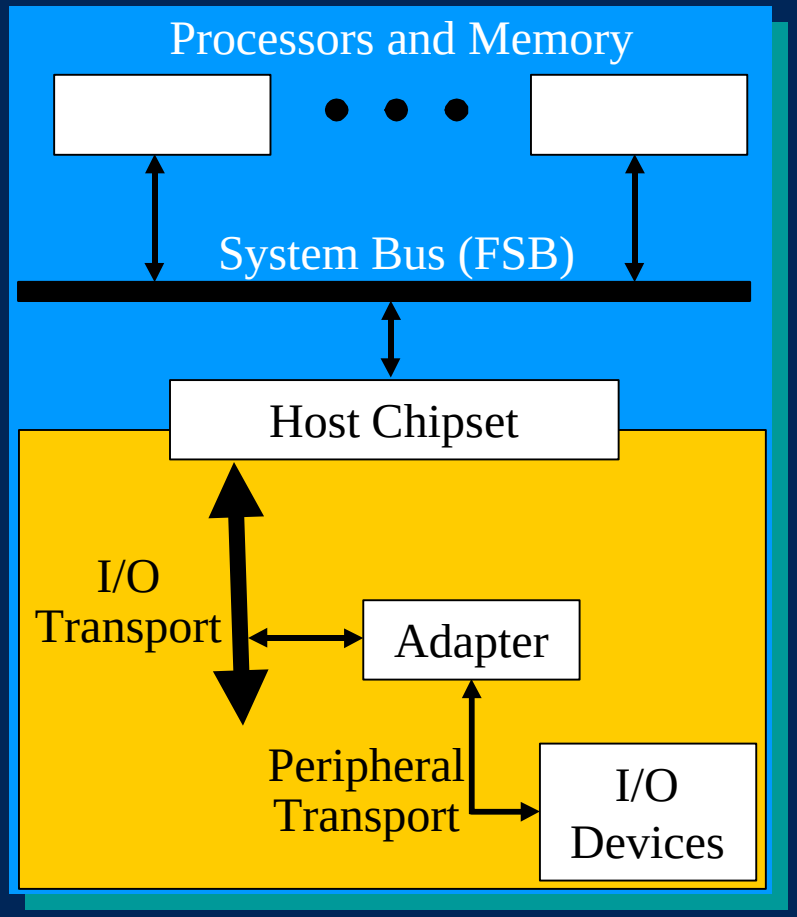


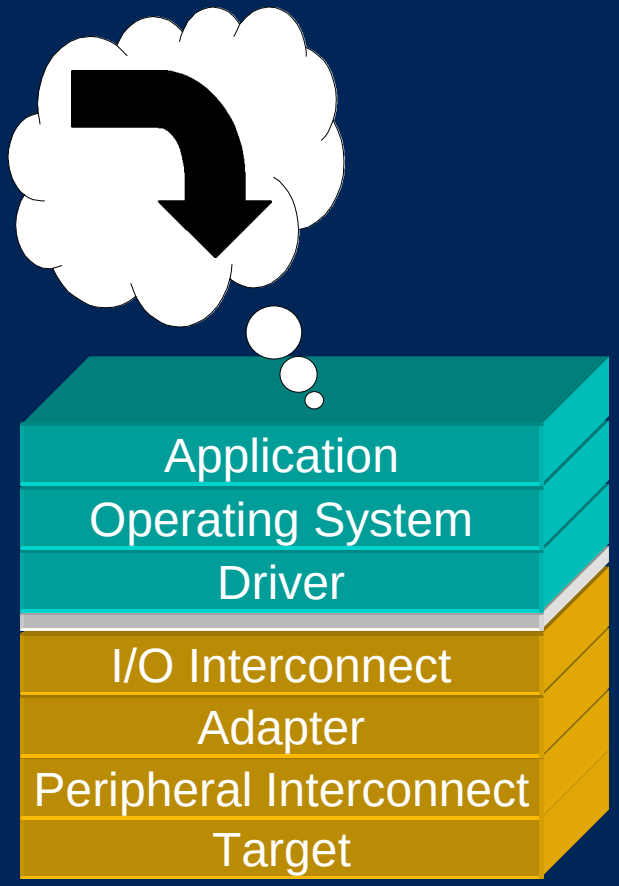
- Define the system (NT servers!)
- Identify the “solution stack”
- Characterize I/O traffic through the stack
- Define important metrics
- Measure critical stack components
- Analyze measurements
- Postulate/suggest improvements
- Redefine system, refine process…repeat!

Process applies equally to storage, network, or IPC

The Science of Measurement

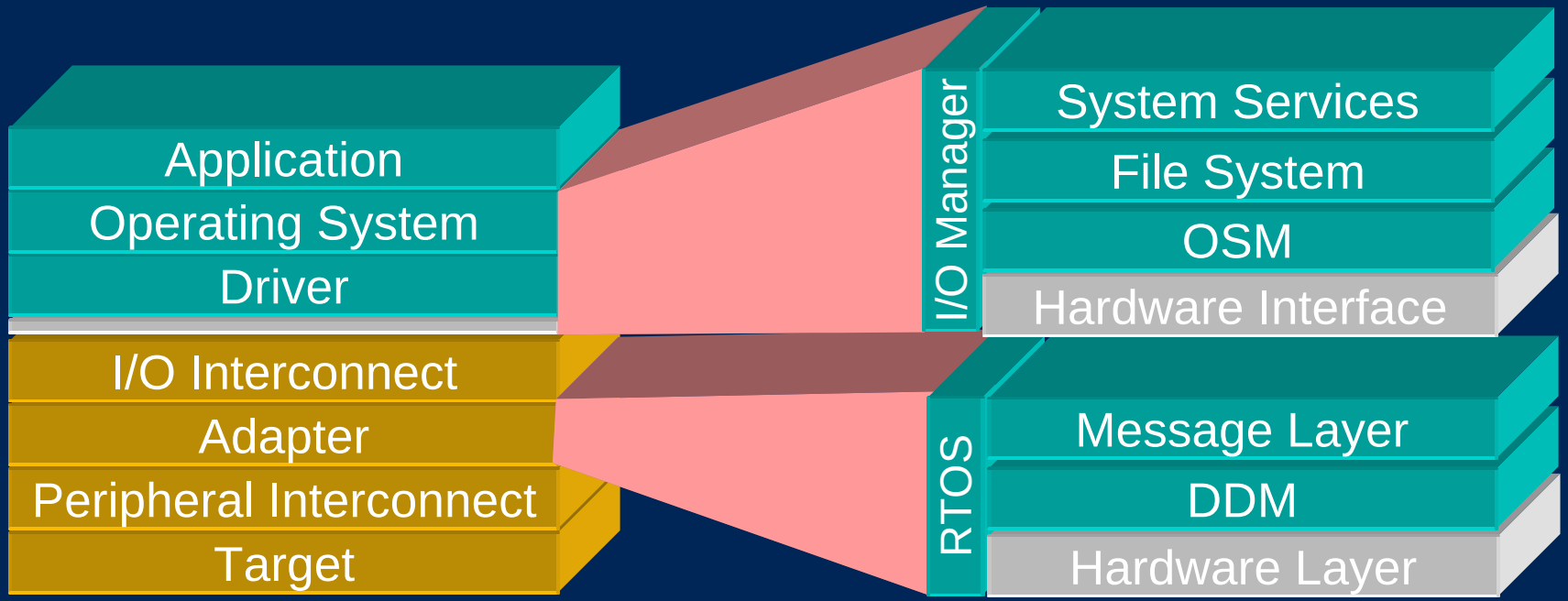
Abstract system into solution stack





The Science of Measurement

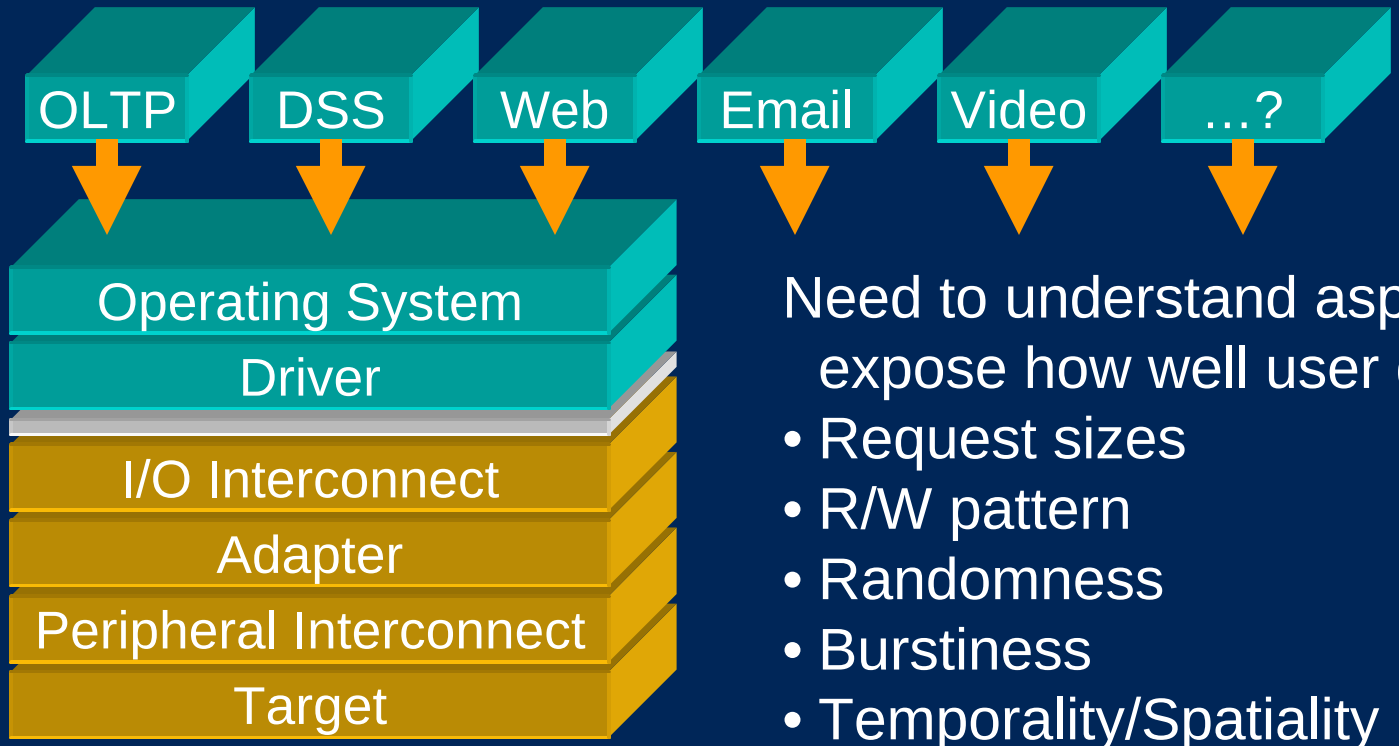
Goal is to find bottlenecks



- Focus on I/O subsystem
- Expand stack as necessary
- Measure accessible components
- Public transports and shared resources often become bottlenecks, which in turn determine throughput and scalability

I2O stack example shows we gain insight as I/O subsystems become more standardized

The Science of Measurement
I/O Characterization Is Hard



Need to understand aspects that expose how well user data is moved:

- Request sizes
- R/W pattern
- Randomness
- Burstiness
- Temporality/Spatiality
- Other

True characterization takes into account all parts of the system, but we can approximate at the I/O Interconnect boundary

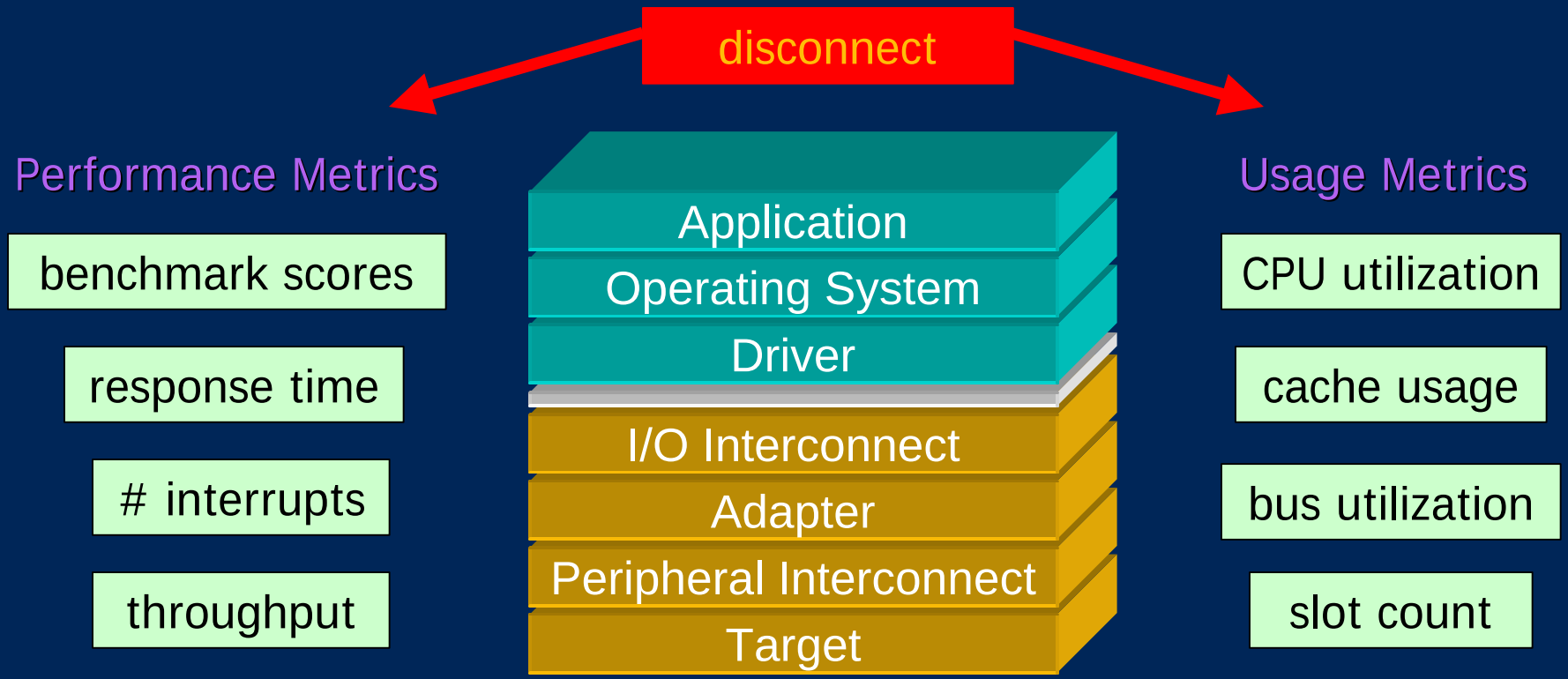
The Science of Measurement

What Metrics Are Important?

- ***Traditional Metrics***
  - Performance Metrics
    throughput, I/O rate, response time, interrupts, benchmark scores, etc.
  - Usage Metrics
    CPU utilization, bus utilization, memory footprint, connectivity, etc.
- ***Enhanced Metrics***
  - Efficiency
  - Effectiveness
  - Transactions per I/O
  - Interrupts per I/O

The Science of Measurement

The problem with *traditional* metrics



Traditional metrics are useful and necessary

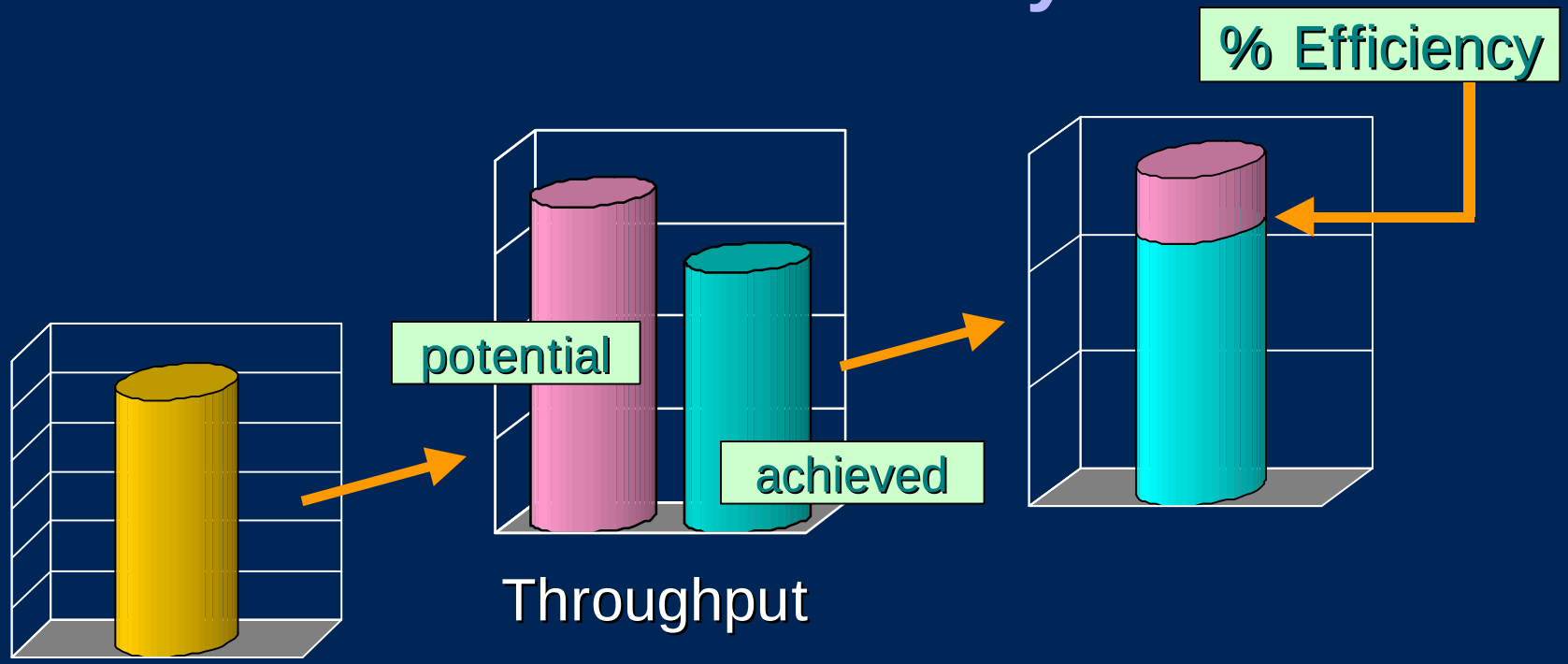
But they may not identify limitations...

What about scalability?

...Or technology comparisons?

The Science of Measurement

Enhanced Metrics: Efficiency



$$\text{Efficiency} = \frac{\textbf{Achieved Throughput}}{\textbf{Potential Throughput}}$$

Efficiency indicates how well application data moves through the stack

The Science of Measurement

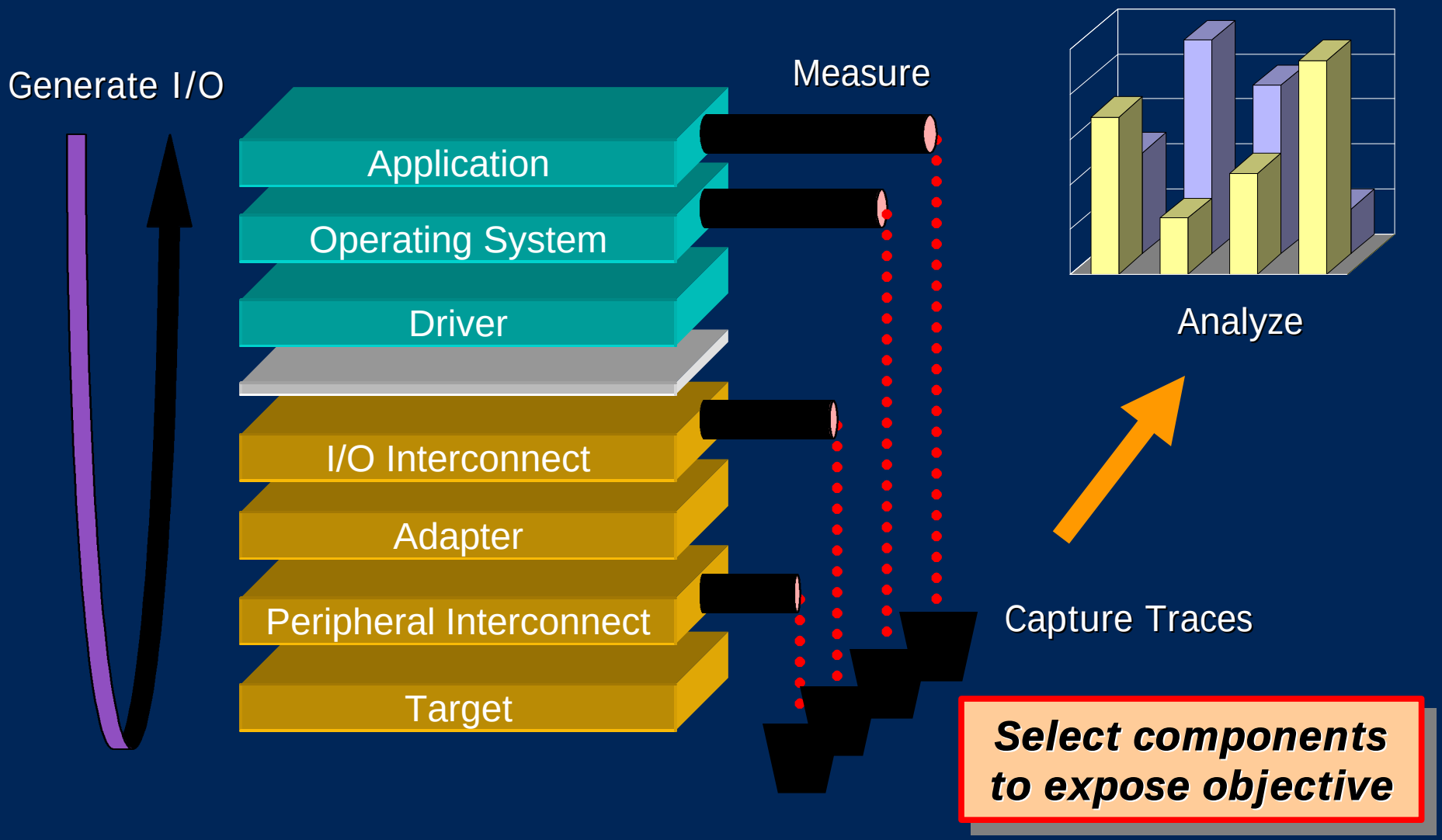
Enhanced Metrics: Effectiveness

Want highest possible throughput with lowest possible overhead

Effectiveness indicates subsystem scaling potential

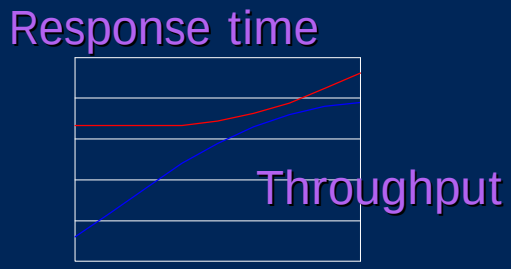
The Science of Measurement

Basic Measurement Techniques

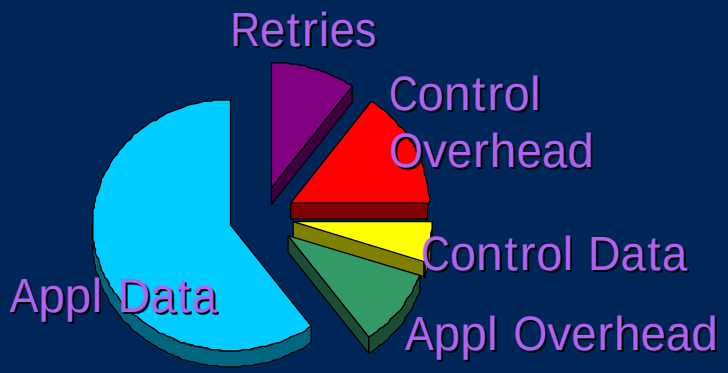


The Science of Measurement

Statistical Analysis

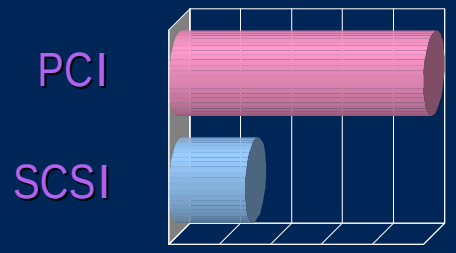


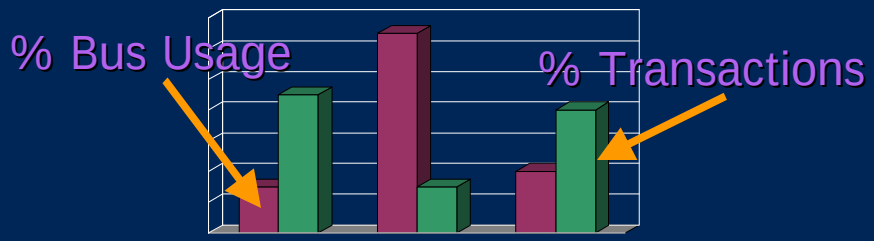
Performance Metrics



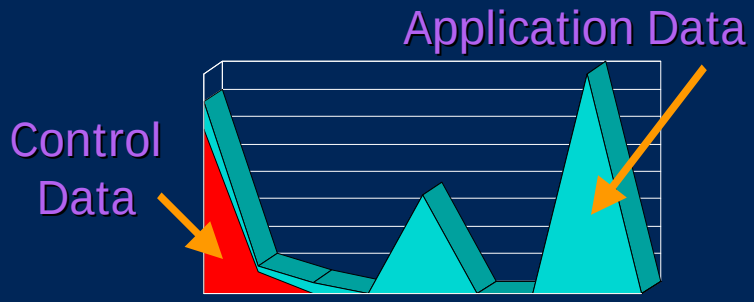
Bus Usage

Transactions Per I/O





Command Usage

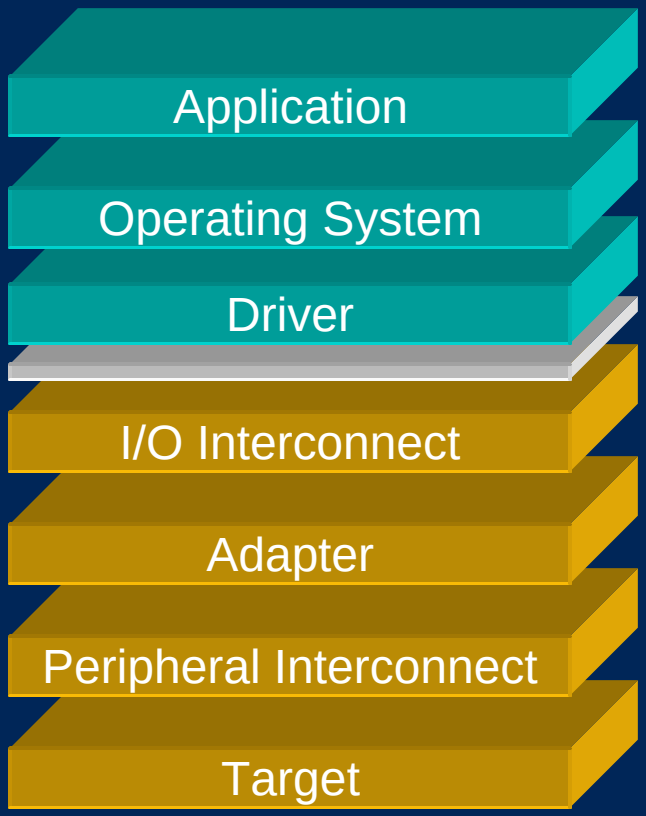


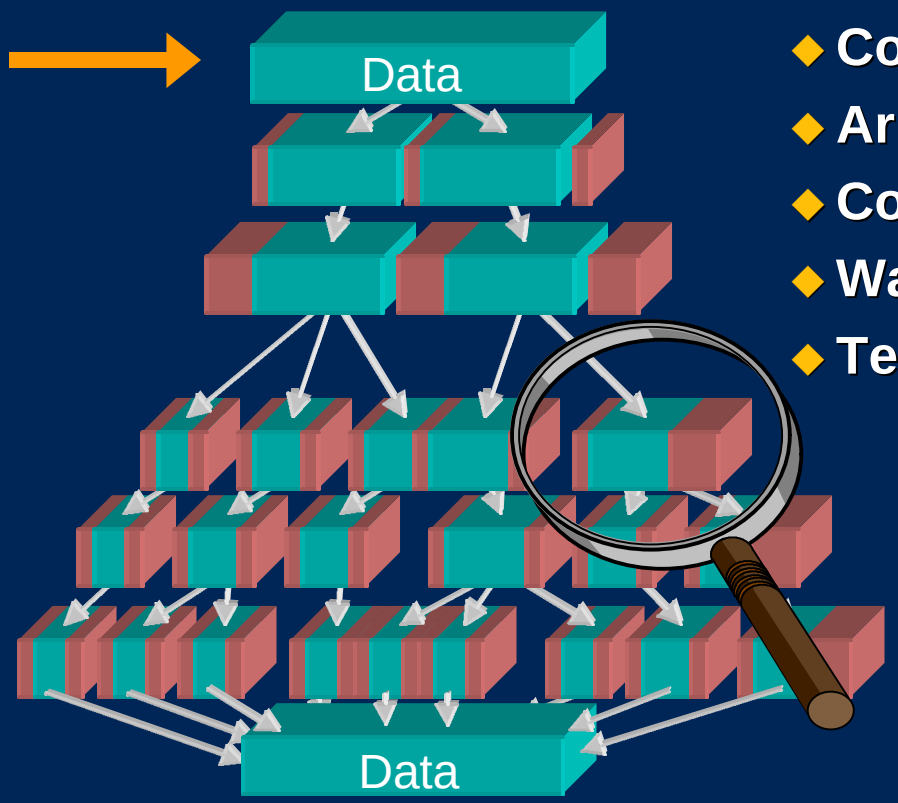
Burst Length

Statistical performance overview highlights problem areas

The Science of Measurement

Transactional Analysis





Overhead

- Setup
- Connection
- Arbitration
- Control data
- Wait states
- Tear down

Transactional analysis clearly points out detrimental behavior

The Science of Measurement

Measurement and Analysis Tools

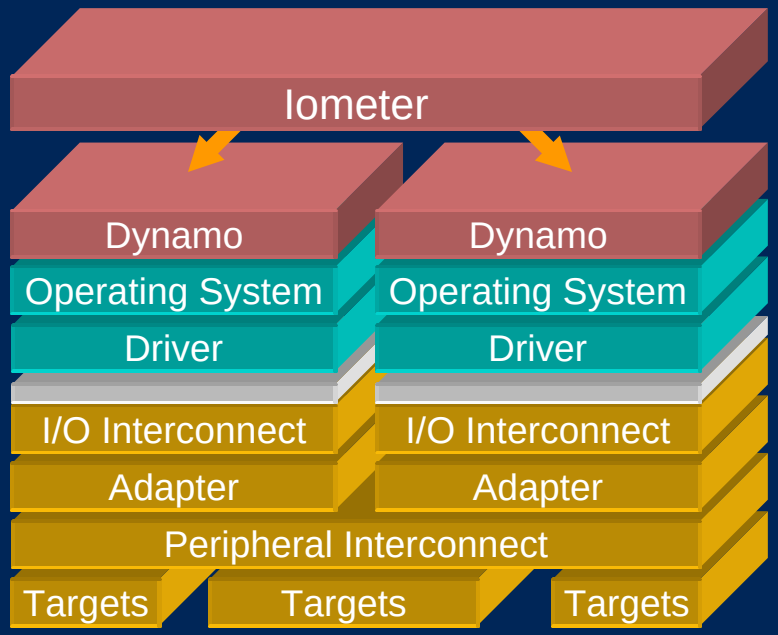
- *Iometer*™ (formerly known as *Galileo*)
  - I/O generation and measurement tool
  - distributed to industry as common framework
- *I/Opener*™
  - *Statistical* and *Transactional* analysis of PCI bus and FC-AL loops using output from PCI and FC analyzers
  - Can be used with Iometer, benchmark, or live workloads

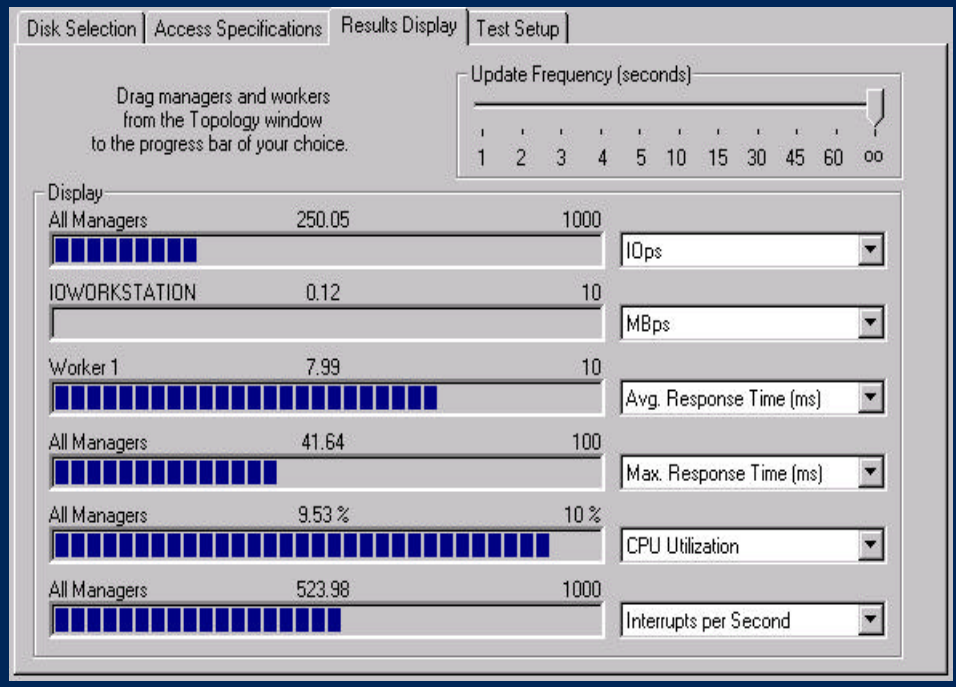
Evolving server I/O subsystem analysis to higher levels of sophistication and automation

The Science of Measurement

Iometer™ Features

- **Gives rapid overview of I/O subsystem characteristics (NT)**
  - ⇨ **flexible I/O workload generator**
  - ⇨ **measures performance at appl level**
  - ⇨ **automated repeatability**
  - ⇨ **low overhead, scalability focus**
  - ⇨ **powerful result wizard**
  - ⇨ **single or clustered systems**

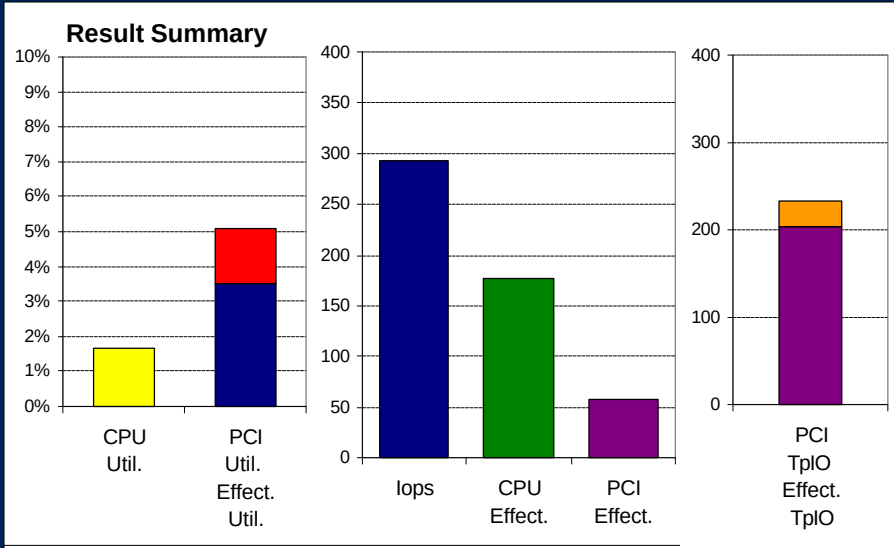


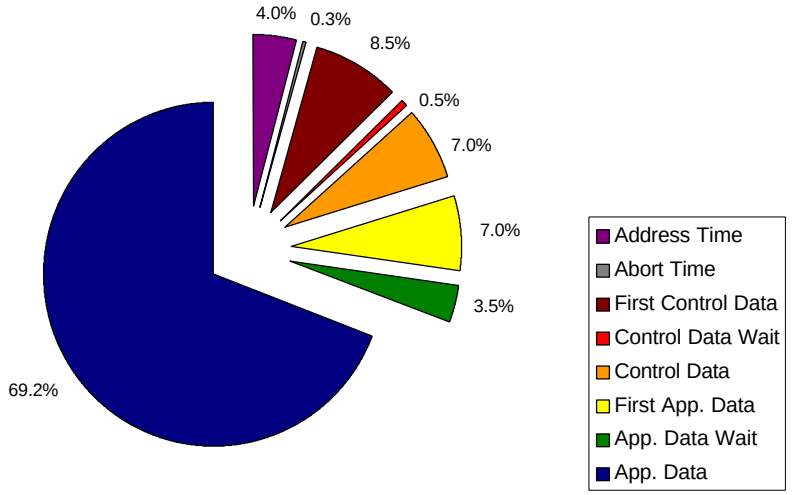


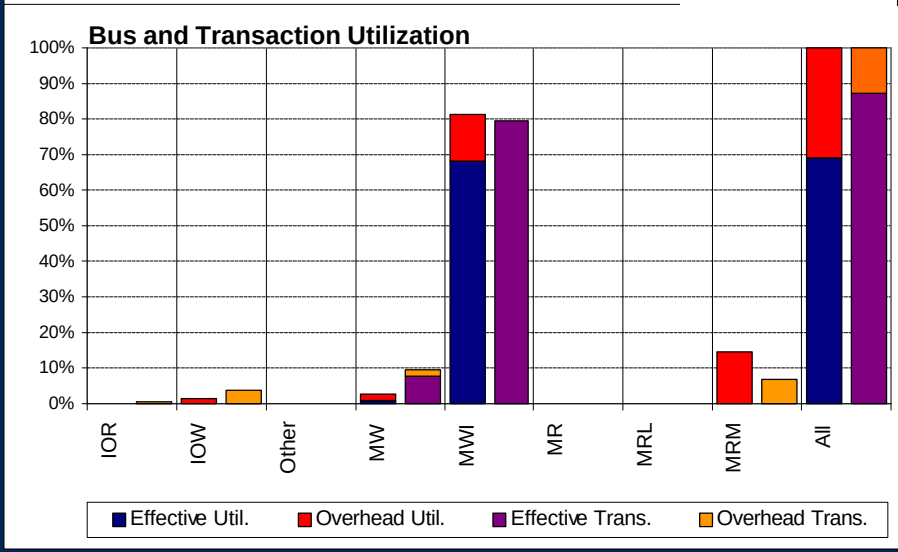
Distributed to the industry; send email to: iometer@intel.com

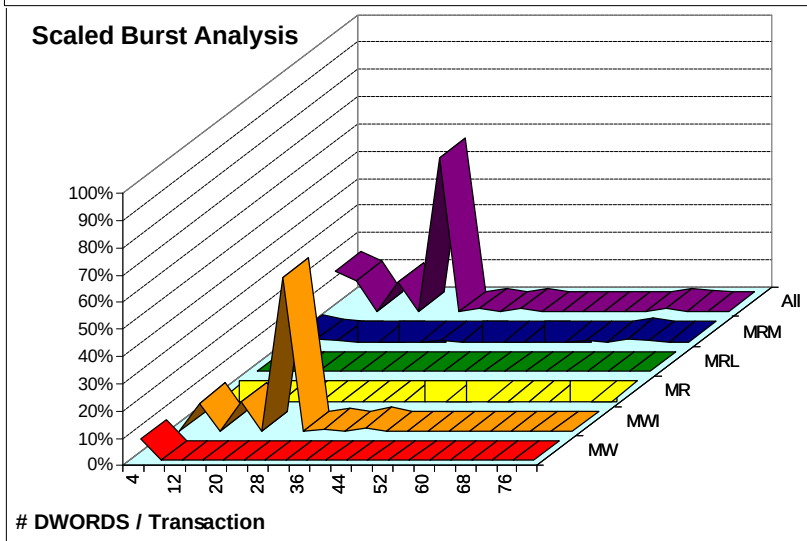
The Science of Measurement

PCI I/Opener™ Tool Sample Output







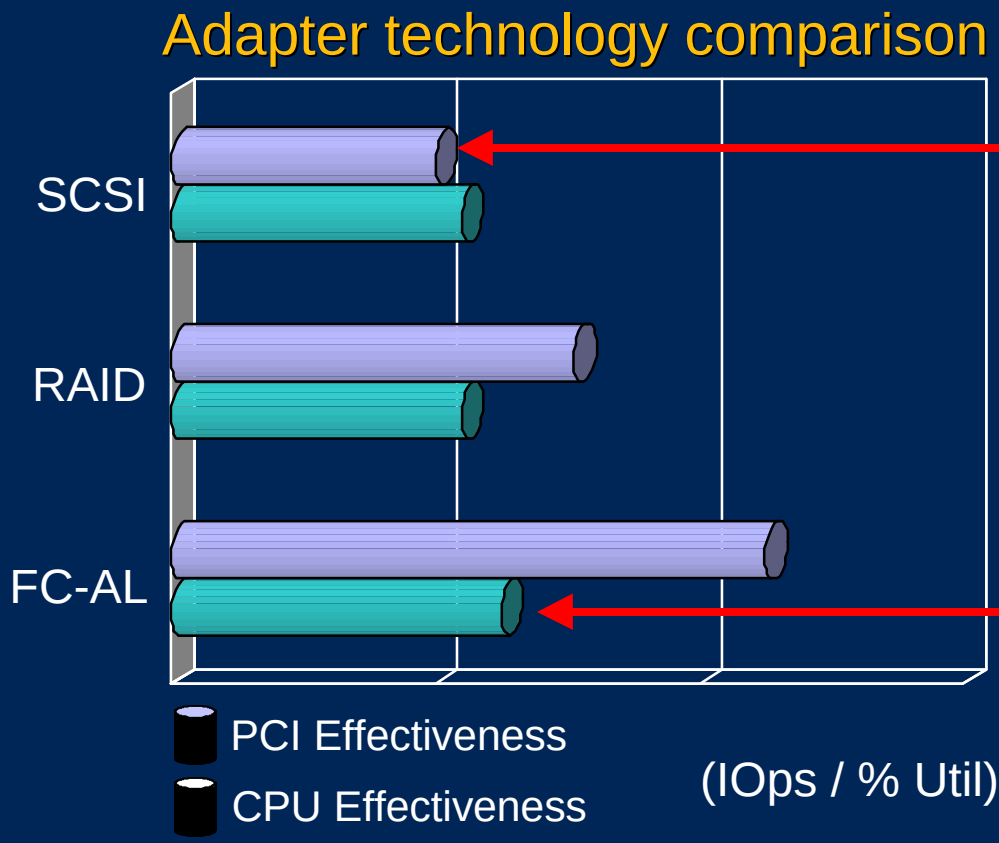


Data from Intel research

Agenda

- Why Measurement and Analysis?
- The Science of Measurement
- Analysis Insights
- Summary

Analysis Insights
Measure I/O Subsystem Scalability

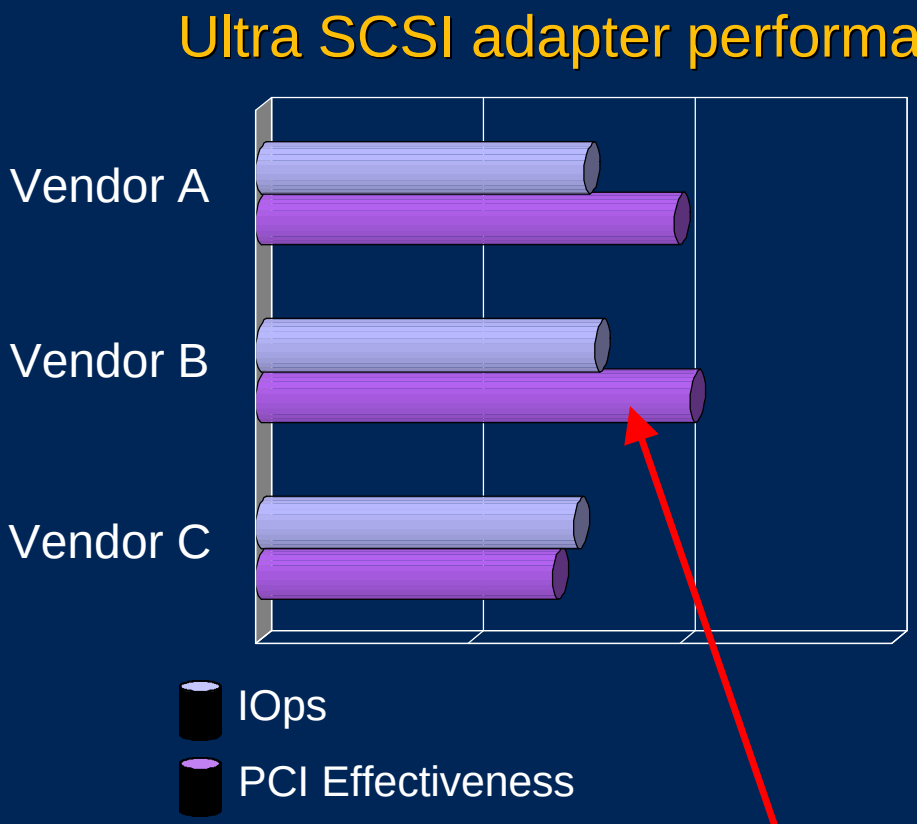


Worse PCI effectiveness means fewer adapters can make use of the PCI bus

Better CPU effectiveness means the system has headroom to potentially handle more adapters

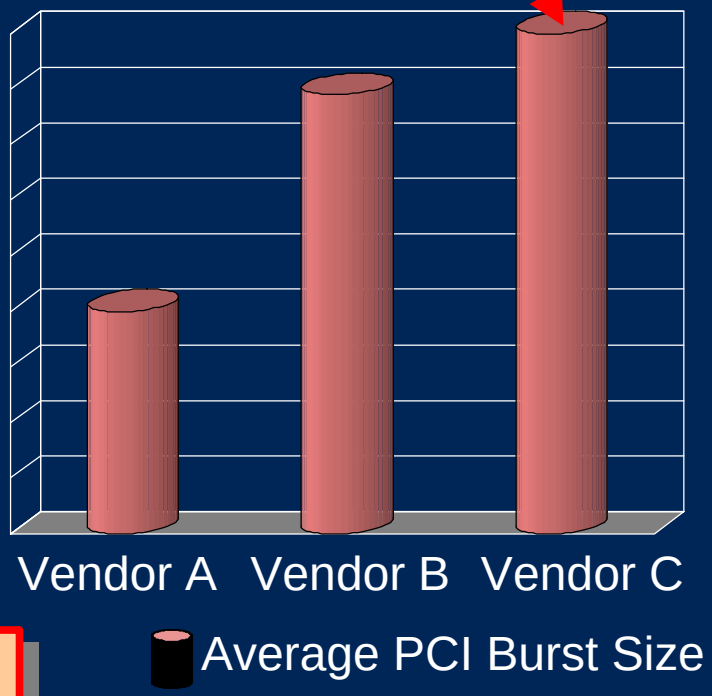
Data from Intel research

Analysis Insights
Compare Adapters



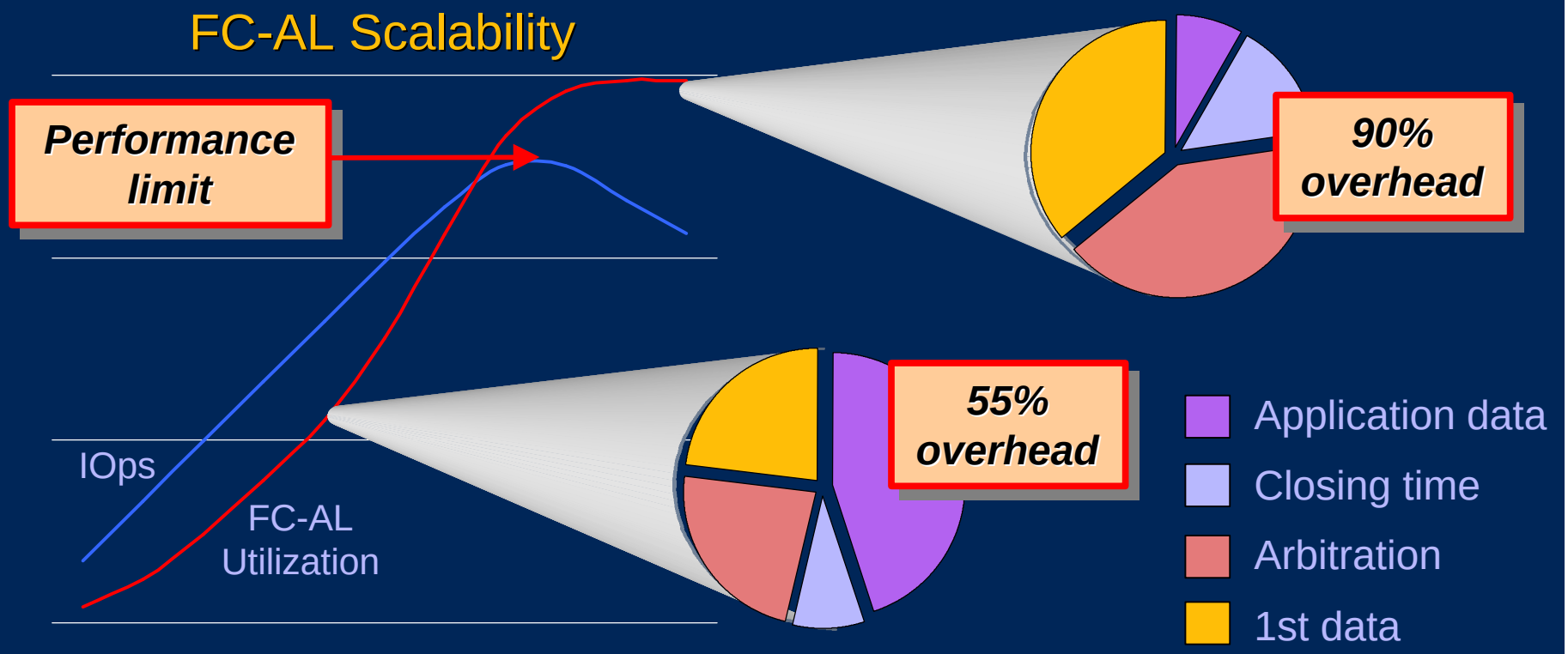
Best throughput and PCI effectiveness

Longer PCI bursts may be better on a shared bus



Data from Intel research

Analysis Insights
Evaluate Emerging Technologies



Determine actual technological advantages

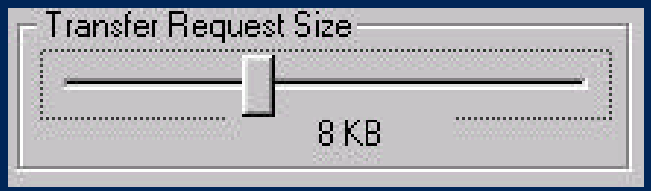
Data from Intel research

Analysis Insights
Eliminate Unnecessary Behavior

Adapter PCI Bus Usage

Address	Command	Bytes Transferred
0x052a0020	4 MWI	512
0x052a0020	65 MWI	8 K

Looking good so far!



But only 8KBytes were requested…!

Driver sent the first 512 bytes twice!

Data from Intel research

Analysis Insights
Eliminate Unnecessary Behavior

Why so many control transactions?

Same data

Address	Command	Data
0x00f66da0	MR	0x00f639b1
0xfe8fdc10	MW	0x00f639b1
	2 MR/MW pairs	
0x00f639c0	MR	
	4 MR/MW pairs	

Adapter PCI Bus Usage

No relevance or repeated control data!

Data from Intel research

Analysis Insights
Improve Driver Code

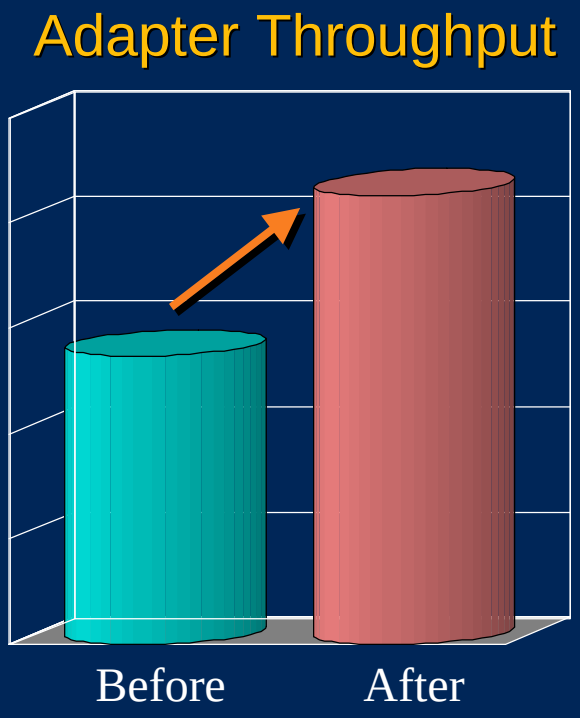
Needs 3K control data for single 8K transfer!

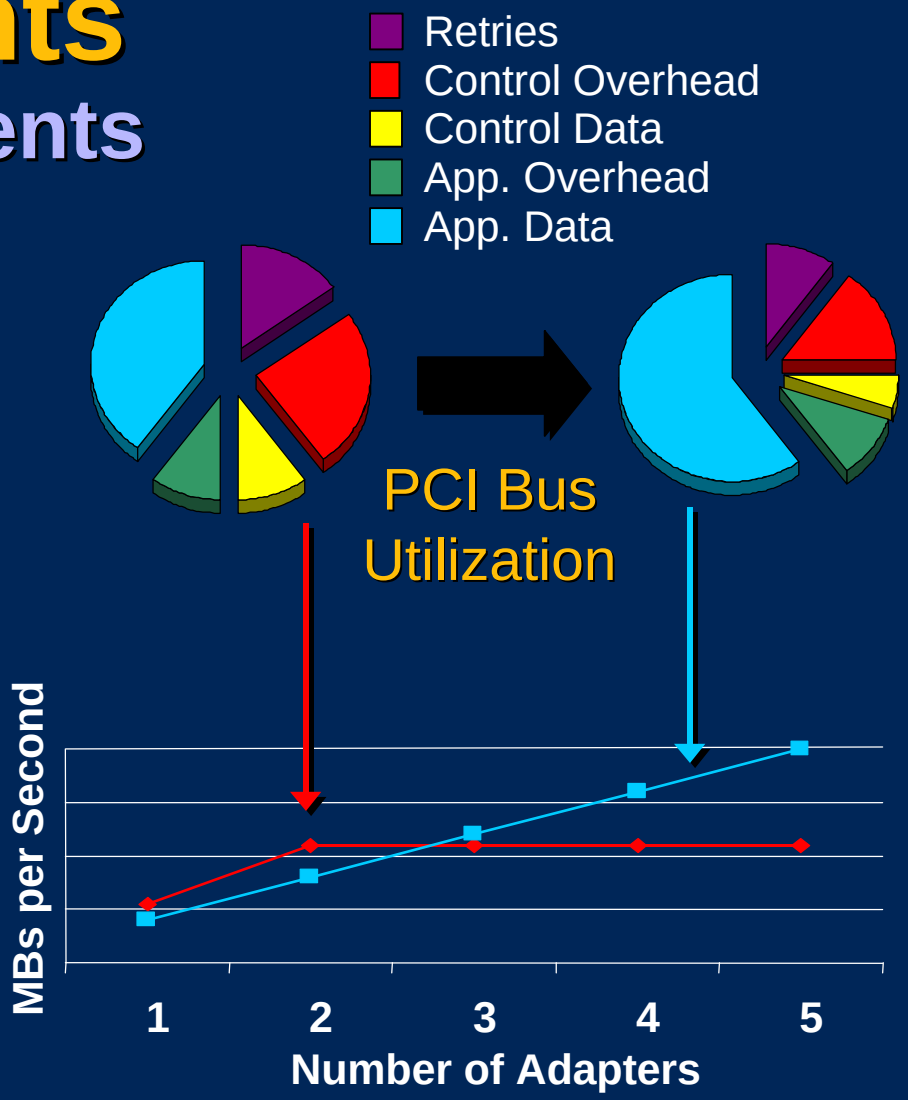
Adapter PCI Bus Usage

Address	Command	Bytes Transferred
0x00f448xx	103 MRM	3196
0x050a0620	***14 MWI***	***2528***
0x0575a000	***29 MWI***	***4128***
0x050a0020	***7 MWI***	***1536***

Data from Intel research

Analysis Insights
Guide I/O Improvements





Where is the most bang for the buck?

Data from Intel research

Agenda

- Why Measurement and Analysis?
- The Science of Measurement
- Analysis Insights
- Summary

Summary

- Continue to push I/O subsystem limits
- Ensure that products use system resources efficiently
- Deeper analysis yields better results
- Enhanced metrics provide additional insight
- Adopt techniques, tools, and terminology to uncover the data that you need
- Real data means real improvements
- See the whole I/O story to ensure high volume NT server success

Call to Action

- Standardize your measurement and analysis techniques
  - Measure scalability
  - Perform competitive analysis
  - Extend and use enhanced system metrics
  - Expand techniques to incorporate industry standards wherever possible

Eliminate guesswork and make informed decisions

Call to Action

- Start using Iometer - available now!
  - Analyze and use results to drive designs
  - Register to receive update notifications

http://developer.intel.com/design/servers/devtools/iometer

iometer@intel.com

- Characterize I/O traffic and build a catalog of important workloads
  - share them with Intel and industry

Legal Disclaimers

1) All data within is derived from research conducted in Intel labs
2) Third-party marks and brands are the property of their respective owners
3) Performance tests and ratings are measured using specific computer systems and/or components and reflect the approximate performance of Intel products as measured by those tests. Any difference in system hardware or software design or configuration may affect actual performance. Buyers should consult other sources of information to evaluate the performance of systems or components they are considering purchasing. For more information on performance tests and on the performance of Intel products, reference intel.com/procs/perf/limits.htm or call (U.S.) 1-800-628-8686 or 1-916-356-3104.